Standardization in Multimodal Content Representation: Some Methodological Issues

Harry Bunt, Laurent Romary

▶ **To cite this version:**

Harry Bunt, Laurent Romary. Standardization in Multimodal Content Representation: Some Methodological Issues. 4th International Conference on Language Resources and Evaluation - LREC'04, 2004, Lisbonne, Portugal. 28 p. inria-00100199

HAL Id: inria-00100199
https://inria.hal.science/inria-00100199

Submitted on 14 Jul 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Standardization in Multimodal Content Representation: Some Methodological Issues

Harry Bunt*, **Laurent Romary**†

* Computational Linguistics& AI, Tilburg University
P.O. Box 90153, 5000 LE Tilburg, Netherlands
harry.bunt@uvt.nl

Laboratoire LORIA
B.P. 239 54506, Nancy, France
laurent.romary@loria.fr

Abstract

In this paper we discuss some basic methodological considerations of the activities undertaken in the ACL-SIGSEM Working Group on the Representation of Multimodal Semantic Information. This independent expert group was founded on the instigation of the International Organization for Standardisation ISO for investigating the possibilities to develop well-founded guidelines for the representation and annotation of semantic information in interactive multimodal contexts, with the aim to support the interoperability and reuse of multimodal and language resources.

1. Introduction

In response to the growing recognition of the importance of interoperability and commonality to enable the sharing, merging and comparing of language resources for developing NLP applications, the International Organisation for Standardisation (ISO) has formed a subcommittee (SC4, Language Resources Management) dedicated to the preparation of international standards and guidelines for the representation and annotation of linguistic data. Proposals are for example under development for morphosyntactic annotation (Clément & De la Clergerie, 2003), for lexical representation (George, 2003) and for an XML-based representation format for feature structures (Lee, 2004).

Another potentially important area was recognised to be that of representing the semantic content of linguistic data and, more generally, of multimodal data, where language is used in combination with other modalities. As a first step in this area, an independent expert group was formed in 2002 within ACL SIGSEM, the Working Group on the Representation of Multimodal Semantic Information (MMSemR). The activities of this Working Group so far consist primarily of:

1. studying and comparing existing representational systems and their underlying principles (see e.g. Bunt, 2003);
2. identifying commonalities in different approaches to semantic representation and annotation;
3. developing methodological principles for identifying and characterising representational concepts for multimodal content.

The present paper discusses some initial steps relating to (3), inspired by ongoing discussions in the MMSemR Working Group and by methodological considerations that have emerged from ISO activities in other domains.

It may be noted that, in general, ISO activities in the area of language technology are not so much aimed at standardization in the sense of proposing particular formats that should be used, but rather at identifying and documenting valuable and common concepts in different approaches and providing guidelines for using these concepts to support interoperability and reuse of resources.

2. Semantic Content Representation

The representation of semantic content is crucial for intelligent multimodal dialogue systems, where users may for example combine speech with graphics, gestures, and the use of facial expressions and where the system may combine speech, text, graphics, and other sounds and visual elements. An intelligent multimodal interface requires the fusion and coordination of multimodal input and output at a semantic level, taking into account the current state of the interaction and the context. The communication between the components of such a system relies on an enabling representation system that should support all stages of multimodal input processing and output generation.

Multimodal interactive systems combine more than one 'language' in which to exchange information, typically including some form of natural language (NL) in combination with graphics, i.e. with a gestural language. This means that multimodal semantic content includes NL semantic content. In view of the overwhelming amount of ambiguity and vagueness inherent in NL, computational semantics has in recent years moved in the direction of computing underspecified representations, that capture semantic information in an utterance rather than represent 'the meaning' of the utterance. This is a useful development for multimodal semantic fusion, and it has the effect that the difference between semantic representations and semantic annotations becomes gradual rather than principled, thus allowing the use of common concepts.

Prerequisite to semantic information representation is a well-delineated notion of 'meaning' which is appropriate in a multimodal context. Bunt & Romary (2002) have pro-

posed to define meaning as the way in which an utterance is meant to change the information state of an interpreting system upon understanding. (The final clause of this definition serves to exclude state-changing effects due to processing beyond establishing the utterance meaning). This definition is broad in the sense that it includes aspects of meaning which are often regarded as "pragmatic" rather than semantic; it is also "pragmatic" in flavour in considering *utterance* meaning, i.e. meaning in context, rather than sentence meaning.

In the area of multimodal content representation, ISO does not aim at developing a standard fixed representation *format*. That would be seen by many researchers as a hindrance rather than as support for their activities. Instead, future ISO activities could sensibly aim at providing well-defined concepts ('data categories' in ISO jargon, or 'descriptors') as a basis for semantic representation and annotation. Note that 'descriptors' (or 'data categories') are abstract concepts, not elements of a particular format or representation language.

A semantic descriptor will often make sense only in combination with certain other descriptors, and not in combination with others, reflecting that alternative semantic theories use different concepts. The ultimate goal is therefore not to propose a single set of descriptors to be used, but a larger set from which coherent subsets can be taken according to one's theoretical preferences. If $\{C_1, C_2,.. C_n\}$ is such a coherent set, its use in building representations means that a representation language L_a has corresponding descriptive terms $\{c_{a1}, c_{a2},.. c_{an}\}$.

One of the requirements on any content representation system is that of "semantic adequacy", by which is meant that the representation language itself has a well-defined semantics. For the representation language L_a, a model-theoretic semantics can be given by specifying a model which will have elements $\{d_{a1}, d_{a2},.., d_{an}\}$ as denotations of the corresponding descriptor names in L_a. For instance, a representation in Minimal Recursion Semantics (Copestake et al., 1995) may have the part [AGENT: john], containing the descriptor name AGENT, which in a model $M_a = < D_a, F_a >$ corresponds to a set of pairs of individuals, such that the second element of each pair is the agent of the event corresponding to the first.

3. Models, Data Models, and Metamodels

Looking for commonalities in alternative approaches, as is typical for ISO work, often implies looking at alternative *models*. In particular, when computational resources are concerned, the notion of a *data model* is often used to capture basic aspects of a particular approach, usually in a semi-formal way. Codd (1970) is often credited for having introduced this for discussing the organization and meaning of the contents of data bases (in particular his famous 'relational data model'). If we want to capture what several models have in common, we may move to a more abstract level, and this is where the term 'metamodel' has emerged. The claim of the present paper is that this semi-formal notion of metamodel may be construed more formally by relating it to the notion of model as used in model-theoretic semantics, and may be helpful as a methodological tool for the definition of abstract concepts for the representation and annotation of semantic information.

A model, in model-theoretic semantics, does two things: (1) it provides the basic ingredients from which denotations can be constructed for the terms of a representation language; and (2) it assigns denotations to the descriptive terms of the language. For instance, a model for the language of standard first-order predicate calculus, PL_1, the mother of all representation languages, has the form $M = < D, F >$ where D is a set of individuals and F is a function assigning to descriptive terms of PL_1 either individuals (in the case of individual constants) or sets of k-tuples of individuals (in the case of k-ary predicate constants).

The aim of the MMSemR Working Group to identify commonalities in different approaches to semantic representation and common representational concepts, disregarding representational formats but with a concern for well-defined concepts, calls for a methodological basis in providing a semantics for abstract representational concepts. This must be more abstract than a standard model-theoretic semantics in two respects:

1. it is not concerned with a particular representation *format*, or *language*, hence not with particular descriptive *terms*;
2. it is not concerned with a particular domain of individual objects.

Instead, we are concerned with such considerations as that we want to able to represent information about events and about individuals participating in events, i.e., we are concerned with *descriptive categories*, and *descriptive concepts*. Providing a semantics for descriptive categories and concepts then comes down to two things:

1. the specification of the *semantic categories* corresponding to descriptive categories;
2. the specification of the semantic objects corresponding to specific descriptive concepts.

A metamodel is an abstract specification of the kinds of descriptors that are considered, and how they are model-theoretically supported. In the case of PL_1, a metamodel should say that the descriptors which are considered are (1) descriptors of individuals (*id*s), and (2) descriptors of predicates (*pd*s); their semantic support by a model should be such that individual descriptors correspond to individuals, and k-ary predicate descriptors to sets of k-tuples of individuals. This can be formulated mathematically by defining a *first-order metamodel* as a pair:

(1) $MM^{(1)} = < \mathsf{ind}, \{(id \rightarrow \mathsf{ind}), (pd \rightarrow \mathsf{S}_k[\mathsf{ind}])\} >$

where ind is the type of individuals, and $\mathsf{S}_k[t]$ is the type of sets of k-tuples of elements of type t (with $0 \leq k$). Characteristic for a first-order model is that it supports only representations in terms of individuals and predicates applied to individuals (first-order predicates, expressing properties of and relations between individuals). In particular, it does not support predicates applied to first-order predicates. A second-order metamodel, supporting also second-order predicates, has the following structure:

(2) $MM^{(2)} = < \mathsf{ind},\ \{(id \rightarrow \mathsf{ind}), (pd^{(1)} \rightarrow \mathsf{S}_k[\mathsf{ind}], (pd^{(2)} \rightarrow S_n[S_k[\mathsf{ind}]])\} >$

for $k, n \geq 1$, where $S_n[S_k[D]]$ is the set of n-tuples of elements of $S_k[D]$, i.e., the set of properties of (and n-ary relations between) properties of (and relations between) individuals.

Metamodels can be seen as characterizations of a class of models, and may be useful for indicating the general approach to a certain task involving complex information. As an example, consider the following two alternative views on feature structures in linguistics.

The first is that of feature structure as a *graph* viewed as a set-theoretical construct. Carpenter (1992), for example, defines a typed feature structure as, given a set Feat of features and a set Type of (hierarchically ordered) types, a quadruple

(3) $< N, n_0, \theta, \mathcal{F} >$

where N is a finite set whose elements are called *nodes*; where $n_0 \in N$, where θ is a total function from N to Type (typing) and where $\mathcal{F}$ is a partial function from $N \times \mathsf{Feat}$ to N (defining arcs, labelled with feature names, that connect the nodes). The node n_0 is the root of the graph; every node in N is required to be reachable from the root node. Pollard and Sag (1987) use this view when they introduce feature structures as semantic entities in the interpretation of representations of linguistic information. They refer to graphs as "modelling structures", i.e., as structures that play a role in models, and they introduce AVMs as structures in a "description language" that is to be interpreted in terms of feature structures-as-graphs: *"Throughout this volume we will describe feature structures using attribute-value (AVM) diagrams"*. (Pollard & Sag, 1987, 19–20).

This view corresponds to the following metamodel that distinguishes nonterminal and terminal nodes and types:

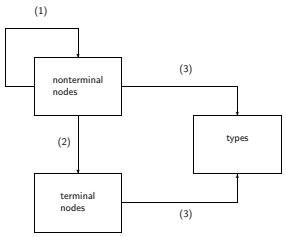


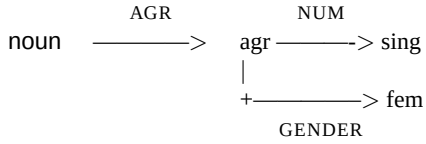
Diagram 1: Metamodel with graphs as model elements

Relations of type (1) in this metamodel correspond to features like HEAD-DAUGHTER in HPSG, those of type (2) to atomic-valued features like GENDER, and those of type (3) to the typing function θ.

An alternative view is that of graphs as *representations*, as a notational alternative to AVMs rather than as the objects interpreting AVMs. For example, Lee (2004) introduces feature structures as ways of capturing information, and mentions graphs as a *notation* for feature structures. Aware of these alternative possible views, Pollard & Sag (1987) note that *"A common source of confusion is that feature structures themselves can be used as descriptions of other feature structures."* One way to avoid confusion is to consider the metamodels corresponding to alternative views.

In the graphs-as-representations view, the graph (4) and the AVM (5) are seen as equivalent representations that can both be interpreted as representing the complex predicate (6).

(4)



(5) $\begin{bmatrix} \textit{noun} \\ \text{AGR} \begin{bmatrix} \text{NUM sing} \\ \text{GENDER fem} \end{bmatrix} \end{bmatrix}$

(6) $\lambda x : noun(x) \wedge num(x) = sing \wedge gender(x) = fem$

(simplifying slightly). This interpretation reflects a similar view on information as that of first-order logic, with two kinds of individuals: the kind of things that x stands for (words and phrases) and the kind of atomic attribute values like 'fem' and 'sing'. These values are associated with word-like individuals through two-place predicates that are in fact functions; moreover, types such as 'noun' correspond to unary predicates. This corresponds to the metamodel visualized in Diagram 2.

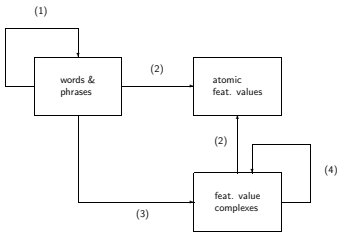


Diagram 2: First-order metamodel for feature structures

Relations of type (1) in this diagram (1) correspond again to features like HEAD-DAUGHTER; (2) to atomic-valued features like GENDER; (3) to features like SYNSEM, and (4) to features like AGR(EEMENT).

4. Metamodels for Semantic Representation

The definition of multimodal meaning mentioned above, in terms of the way in which a communicative action changes the information state of an addressee, has some immediate implications on the basic ingredients that should be available for the description of semantic information. A communicative action is, first of all, an event: something that happens at some point in time (or between two time points). Second, it has at least two participants: the agent who performs the action and the one at whom the action is addressed. Third, these participants have different roles; that of agent and addressee. So we minimally want to distinguish events; entities participating in events; temporal objects for anchoring events; roles relating events to participants, and temporal functions for anchoring events in time.

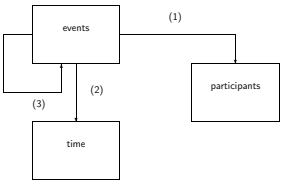


Diagram 3: Metamodel with events, participants, and time

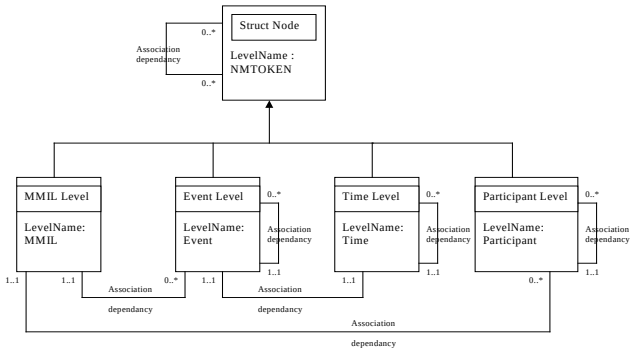


Figure 1: The MMIL Metamodel

This corresponds to the metamodel depicted in Diagram 3.

Events are needed not only for describing communicative events, but also for the semantic analysis of NL sentences, where 'roles' correspond to semantic relations. As such events may have other events as 'participants', we have included functional relations between events in the metamodel of Diagram 3; 'participants' in this diagram should then be understood as 'participants other than events'.

The metamodel shown here is evidently very general, and as such would seem to be of little use. Still, it may be interesting to note that a multimodal interactive system, currently under development jointly by research groups at LORIA in Nancy (France) and the university in Saarbruecken (Germany), the MIAMM system, makes use of a metamodel that is only marginally more complex. This metamodel is shown in Figure 1 in the form of a UML diagram. It supports, in addition to the metamodel of Diagram 3, relations like part-whole within participants and within events, and relations between temporal objects (temporal relations). For further discussion see Kumar & Romary (2003).

Whereas metamodels do indeed seem to be a useful tool in describing semantically well-founded concepts for semantic representation and annotation, just as in ordinary model-theoretic semantics the *intended meaning* of a concept cannot be captured by indicating formal relations with entities in a metamodel. An much needed additional tool is that of a *data category registry*, i.e. a software repository documenting common and valuable concepts (Romary, 2004). At LORIA in Nancy preliminary work has been undertaken to set up such a registry, from which Figure 2 shows a simplified example.

Another interesting application of metamodels to the comparative study of meaning representation is that of characterizing structural models of dialogue, like Rhetorical Structure Theory (Mann & Thompson, 1988), versus information-state change approaches, like Dynamic Interpretation Theory (Bunt, 2000). Unfortunately, limitations of space prevent us from discussing this application in the present paper.

`/Addressee/`
Definition: An entity that is an intended recipient of of a communicative event, whose state of information is intended to be influenced by the event.
Source: (implicit) an event, whose eventType should be `/CommunicativeAct/`. Target: a participant (user or system).

Figure 2: Sample data category registry entry.

5. Concluding Remarks

In this paper we have discussed the notion of a metamodel in relation to data models and models in model-theoretic semantics, and outlined its possible use together with a data category registry in the definition of abstract elementary descriptors of semantic information.

6. References

Bunt, H. (2000) Dialogue Pragmatics and Context Specification. In H. Bunt & W. Black (eds.) *Abduction, Belief and Context in Dialogue.* Benjamins, Amsterdam, 81–150.

Bunt, H. (2003) Underspecification in Semantic Representations: Which Techniques for What Purpose? In *Proc. 5th Intern. Workshop on Computational Semantics (IWCS-5)*, Tilburg (Netherlands), January 2003, 37–54.

Bunt, H. & L. Romary (2002) Towards multimodal content representation. Proceedings LREC 2002 Workshop.

Carpenter, R. (1992) *The Logic of Typed Feature Structures.* Cambridge University Press.

Clément, L., and E. de la Clergerie (2003) Language Resource Management – Morphosyntactic Annotation Framework. ISO Document TC 37/SC 4 N47, KORTERM, Taejon (Korea).

Codd, E.F. (1970) A relational model for large shared data banks. *Communications of the ACM* 13(6): 377–387.

Copestake, A., D. Flickinger and I. Sag (1995) Minimal Recursion Semantics. Unpublished manuscript, CSLI, Stanford.

George, M. (2003) Language Resource Management – Lexical Resource Markup Framework (LMF), ISO Document TC 37/SC 4 N088, KORTERM, Taejon (Korea).

Kumar, A. and L. Romary (2003) A Comprehensive Framework for Multimodal Meaning Representation. In *Proc. 5th Intern. Workshop on Computational Semantics (IWCS-5)*, Tilburg (Netherlands), January 2003, 225–251.

Lee, K. (ed.) (2004) Language Resource Management – Feature Structures Part1: Feature Structure Representation. ISO document ISO/TC 37/SC 4 N 033. KORTERM, Taejon (Korea).

Mann, W. and S. Thompson (1988) Rhetorical Structure Theory: Toward a functional theory of text organization. *Text* 8(3): 243–281.

Pollard, C. and I. Sag (1987) *Head-Driven Phrase Structure Grammar*. CSLI and University of Chicago Press.

Romary, L. (ed.) (2004) Specification of data categories and management of a data category for language resources. ISO document ISO/TC 37/SC 3 N 488. KORTERM, Taejon (Korea).